Using Ensemble Decisions and Active Selection to Improve Low-Cost Labeling for Multi-View Data

Umaa Rebbapragada UMAA.REBBAPRAGADA@JPL.NASA.GOV
Kiri L. Wagstaff KIRI.WAGSTAFF@JPL.NASA.GOV
Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA, 91109, U.S.

Abstract

This paper seeks to improve low-cost labeling in terms of training set reliability (the fraction of correctly labeled training items) and test set performance for multi-view learning methods. Co-training is a popular multi-view learning method that combines high-confidence example selection with low-cost (self) labeling. However, co-training with certain base learning algorithms significantly reduces training set reliability, causing an associated drop in prediction accuracy. We propose the use of ensemble labeling to improve reliability in such cases. We also discuss and show promising results on combining low-cost ensemble labeling with active (low-confidence) example selection. We unify these example selection and labeling strategies under *collaborative learning*, a family of techniques for multi-view learning that we are developing for distributed, sensor-network environments.

1. Introduction

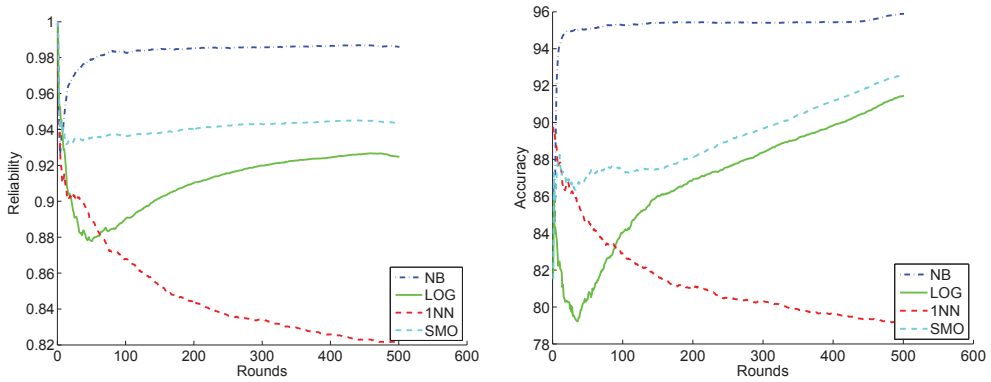
Because supervised learning is constrained by the availability of labeled data, there is extensive research effort on ways to reduce the cost of training data generation. Two such methods are active learning (Cohn et al., 1994) and co-training (Blum & Mitchell, 1998). Active learning aims to make efficient use of a presumably high-cost oracle to label the most informative examples, such as those for which the classifier has low confidence (Lewis & Gale, 1994). Co-training is a seminal semi-supervised learning method (Chapelle et al., 2010) in which each classifier self-labels the data and then shares those for which it is most confident with its neighbor. Thus, co-training combines low-cost labeling with high-confidence example selection, while active learning combines high-cost (oracle) labeling with low-confidence example selection.

Oracle labeling by definition produces labels that are noise-free, while labels produced by an automated classifier will be less reliable. The issue of noisy co-training labels is one that has not received a lot of attention, despite its implications for systems that employ co-training. Pierce and Cardie observed that co-training achieved an initial improvement in accuracy followed by a subsequent decline for a noun phrase bracketing task (Pierce & Cardie, 2001), due to "degradation in the quality of the labeled data." As we would expect, noise in the labels leads to poor generalization performance. Pierce and Cardie solved the label noise problem by employing a (high-cost) oracle to label the selected examples. Obviously, this solution may be unacceptably expensive or infeasible in many settings.

In this paper, we investigate three questions. The first is: *under what conditions does co-training generate unreliable labels?* While Pierce and Cardie demonstrated a case in which co-training produced unreliable labels, there are certainly cases where this is not true. We have identified the base learner's robustness to label noise as an important, previously unidentified factor in co-training success.

The second question is: *how can we improve the reliability of low-cost labeling for multi-view learning?* We propose a solution that assumes the labelers can communicate by querying for and sharing labels. In this setting, we adapt strategies from ensemble labeling which theorize that an ensemble decision is more reliable than an individual decision when each labeler's performance is better than random and the labelers collectively make uncorrelated errors (Dietterich, 2000). To minimize the addition of mislabeled exam-

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.



(a) Reliability (b) Accuracy

Figure 1. Sensitivity of co-training (MC-SLF) performance to the choice of base learner: Naive Bayes (NB), Logistic Regression (LOG), Support Vector Machine (SVM), or 1-Nearest Neighbor (1NN). Left figure shows training set reliability and right figure shows test set accuracy on four-view VLBA data. Results are averaged over 50 runs.

ples into the training set, we also allow the querying labeler to abstain from adopting the ensemble decision when merited.

The third question is: *how can we combine the strengths of active learning and co-training to generate more reliable labels at low cost?* We explore whether it is possible to combine low-confidence (active) example selection with low-cost labeling, and we discuss some preliminary results.

Finally, we unify strategies for low-cost multi-view learning under the umbrella term *collaborative learning* which we are developing for multi-view, distributed sensor networks. Collaborative learning employs active queries and ensemble-based labeling to leverage the information contained in different views. The goal of collaborative learning is to quickly generate large amounts of labeled data from only a few initial labeled items, while continually learning from the data and improving in classification performance.

In this paper, we interleave these questions with illustrative results using data from the Very Long Baseline Array (VLBA) (Napier et al., 1994). We therefore digress briefly to introduce the data set here. This four-view data set consists of simultaneous observations of pulsar B0329+54 from four of the VLBA's ten 25-m radio telescopes; each telescope provides a different view. The observations are aligned across views by their timestamps. We created a classification data set by extracting a short segment of the time series surrounding each known pulse (positive) and periods without a pulse (negative), yielding a balanced data set of 1360 examples. Pulses generally span multiple time steps, so for each event we created a feature vector consisting of 21 values: the flux of the central pulse or non-pulse observation and the 10 preceding and following time steps.

2. When is Co-training Unreliable?

Co-training in its original form trains a pair of Naive Bayes classifiers on independent "views", or representations, of the same (binary) initial labeled data set ($\mathcal{L}$) (Blum & Mitchell, 1998). Learning proceeds as each classifier predicts labels on a random subset of the shared pool of unlabeled examples ($\mathcal{U}$), selects the top p positively and top n negatively labeled examples, and then shares them with the other learner. The values p and n are set proportionally to the underlying distribution of examples. We refer to this example selection strategy as MC, for most confident. A total of $2p+2n$ newly labeled examples are added to the training set, and both classifiers are re-trained on their respective views of $\mathcal{L}$ (Blum & Mitchell, 1998). The process repeats for a user-specified number of rounds. At termination, a final classifier is built by combining the outputs of the individual classifiers.

Although co-training was originally designed for two learners, it is simple to generalize to N learners, one per view. At each round, a learner selects $p+n$ examples, self-labels them, and then shares those newly-labeled examples with all learners. All learners re-train after their training sets are updated. We refer to the co-training learner as MC-SLF because it self-labels examples for which it is most confident.

We implemented the generalized form of co-training (MC-SLF) using Naive Bayes (NB), Logistic Regression (LOG), Support Vector Machine (SVM), and 1-Nearest Neighbor (1NN) classifiers, then ran experi-

ments on the four-view VLBA data (one learner per view). In each experiment, we reserved 25% of the data for testing. The remaining 75% was divided between the labeled ($\mathcal{L}$, initialized with one positive and one negative example) and unlabeled ($\mathcal{U}$) pools. We conducted 500 rounds of learning. In each round, items were selected for labeling from 50 items randomly chosen from $\mathcal{U}$. At the end of each round, each learner was evaluated on its view of the test set. Results are averaged across all learners and 50 independent runs.

Figure 1 reports MC-SLF results on training set reliability (the fraction of correctly labeled items) and test set accuracy. We observe that the choice of base classifier strongly impacts the performance of co-training. All methods start with a training set reliability of 1.0 (Figure 1(a)). Reliability for NB and SVM dips sharply for a few rounds, then rebounds and stabilizes near 98% and 94% respectively. LOG exhibits an early, steep decline in reliability, but recovers to 92%. For 1NN, reliability degrades monotonically to 82%.

Figure 1(b) shows that classifier accuracy degrades along with training set reliability. Co-training has a *negative* impact on learning for 1NN; test set accuracy declines from 90% to 78%. In contrast, NB and SVM benefit from co-training, with only 2 and 6% noise in the training set, respectively. Likewise, LOG exhibits learning after recovering from the initial drop in training set reliability.

The sensitivity of co-training to the choice of base learner highlights the importance of understanding both the base learner's performance on the problem domain and its general sensitivity to noisy labels. Curiously, the field lacks a general consensus on how tolerant various learning algorithms are to label noise. Several studies have tested noise sensitivity empirically (Kalapanidas et al., 2003; Pechenizkiy et al., 2006; Atla et al., 2011), with conflicting conclusions. It may well be that no one algorithm is most tolerant to label noise (cf. No Free Lunch Theorem (Wolpert, 1996)). NB was most robust for this VLBA data, but it failed on the noun phrase bracketing task described above (Pierce & Cardie, 2001).

3. Improving Label Reliability with Ensemble Labeling

Most of the work on addressing the challenges of label noise seeks to remove mislabeled items via preprocessing (Brodley & Friedl, 1999). Methods for improving noise tolerance include decision tree pruning and ensemble methods. The latter helps reduce the influence of individual mislabeled items by combining the votes of multiple classifiers (Dietterich, 2000).

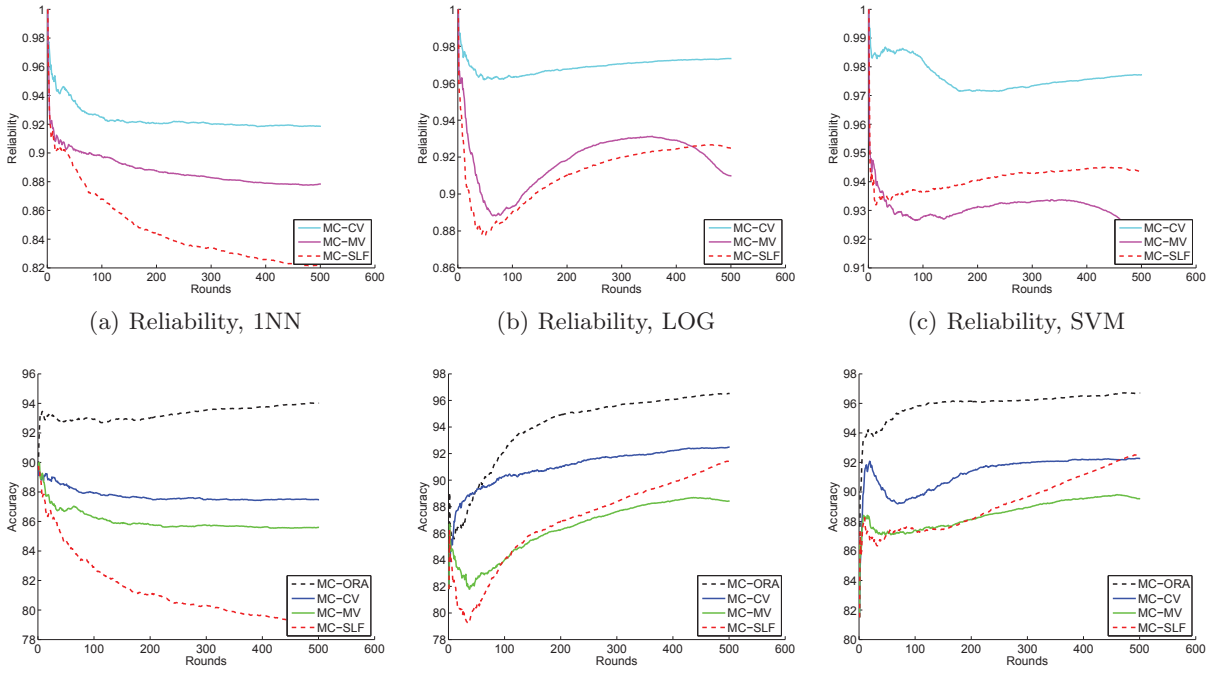
We propose the use of ensemble labeling in the multi-view setting to improve label quality. Co-training requires a networked setup so that self-labeled examples can be shared. Ensemble labeling enables each learner to query other members of the network (neighbors $\mathcal{N}$) for predictions on an example of its choice. The learner combines $\mathcal{N}$'s predictions into a consensus label and adds the item to $\mathcal{L}$. Effectively, each learner has access to its own ensemble for labeling purposes. Our hypothesis is that ensemble labeling will produce a more robust prediction when self-labeling is unreliable.

A labeling ensemble can still generate unreliable labels. Therefore, we also permit each learner to *abstain* from adding an example to its labeled set if the ensemble predictions cannot be unified with high confidence. If a learner chooses to abstain, the example remains unlabeled, the learner is not re-trained, and the process continues with the selection of another example.

We evaluated two naive ensemble labeling strategies with implied abstention policies. The majority vote (MV) policy abstains unless there is agreement on the item's label by a majority of $\mathcal{N}$. The consensus vote (CV) policy abstains unless there is unanimous agreement on the item's label. Both strategies have been used to detect potentially mislabeled items and filter them from the training set in regular supervised learning (Brodley & Friedl, 1999).

Pairing the MC example selection strategy with the MV and CV labeling strategies produces MC-MV and MC-CV. Figure 2 compares co-training (MC-SLF) with MC-MV and MC-CV in terms of reliability and test set accuracy for the three classifiers from Figure 1 on which MC-SLF shows limited or poor performance (1NN, LOG, SVM). As an upper bound on test set accuracy, we also report results for MC-ORA, which is co-training with an oracle labeler (reliability for MC-ORA is always 1.0). MC-CV, which requires a full consensus to generate a new label, attains the best label reliability and the best accuracy for all classifiers. Using an SVM, MC-SLF eventually achieve the same accuracy but at a slower rate. This is consistent with our observation that the SVM is more resistant to label noise than LOG or 1NN. MC-MV out-performs MC-SLF in terms of accuracy and reliability with 1NN, the least robust classifier, and performs equivalently to MC-SLF otherwise.

MC-CV also ends up with a smaller labeled set than MC-MV or MC-SLF, because it requires a consensus. Table 1 shows the size of the labeled training set $\mathcal{L}$, averaged across all views and experimental runs. With-



(a) Reliability, 1NN (b) Reliability, LOG (c) Reliability, SVM

(d) Accuracy, 1NN (e) Accuracy, LOG (f) Accuracy, SVM

Figure 2. Ensemble labeling improves performance over self-labeling on learning algorithms where co-training performs poorly. We compare MC-MV and MC-CV to self- (MC-SLF) and oracle (MC-ORA) labeling in terms of training set reliability (top) and test set accuracy (bottom) using 1NN, LOG, and SVM.

Classifier	Average labeled set size		
	MC-SLF	MC-MV	MC-CV
1NN	1002.00	976.92	446.06
NB	995.78	809.42	513.40
LOG	994.00	942.56	600.46
SVM	978.30	961.32	648.42

Table 1. Size of labeled set after 500 rounds of learning, averaged across 50 runs.

out abstention, the size of $\mathcal{L}$ would be 1002 (because $\mathcal{L}$ starts with 2 examples, one for each class, and receives 2 examples for 500 rounds). However, $|\mathcal{L}|$ can be lower if, in a given round, the learner predicts the same class for all examples (so it cannot select an example from each class). Because CV requires a unanimous prediction, it abstains much more often than MV. However, the items that are added are of higher reliability. In this case, the increased reliability yields better performance than more frequent re-training with less reliable labels (MC-SLF, MC-MV). In experimental results omitted for space reasons, the MV strategy outperforms the CV strategy as the number of learners in the network increases (and consensus is rarer). In this setting, MV produces larger amounts of reliable training data and achieves higher accuracy compared to CV labeling.

4. Accelerating Learning with Active Example Selection

We now examine how active example selection can be applied in a low-cost learning setting. For active selection, we select items with low confidences. This is akin to some active learning strategies, with the key difference that we do not employ an oracle to label the selected items. Confidence is usually measured in terms of posterior probabilities that are either naturally output by the classifier or calibrated from the classifier output (Niculescu-Mizil & Caruana, 2005). We formalize the MC and LC selection strategies as follows. Given $P(l|x)$ as the posterior on the learner's prediction of label l for example x, the MC strategy selects x as $\text{argmax}_{x\in\mathcal{U}} \max_l P(l|x)$, and the LC strategy selects $\text{argmin}_{x\in\mathcal{U}} \max_l P(l|x)$.

Figure 3 compares the performance of oracle labeling using both selection strategies. LC-ORA is similar to the multi-view method Co-Testing (Muslea et al., 2006), except that Co-Testing selects "contention points", or examples on which the learners disagrees, rather than those on which one learner is least confident. The LC example selection strategy is reminiscent of uncertainty sampling (Lewis & Gale, 1994) from active learning. MC-ORA corresponds to Cor-

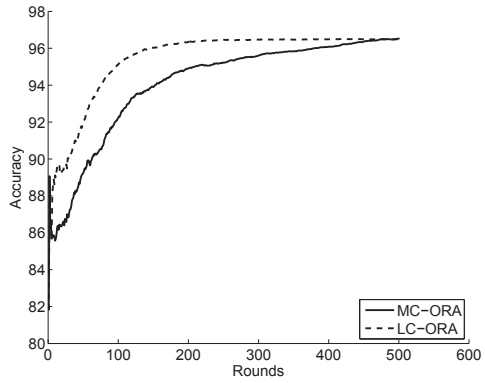


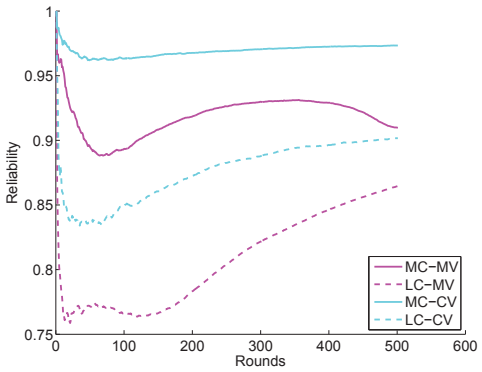
Figure 3. Active example selection (LC) paired with oracle labeling learns more quickly than MC paired with an oracle. Results shown use classifier LOG.

rected Co-training (Pierce & Cardie, 2001). The results show that with high-quality, high-cost labeling, LC-oracle improves performance at a faster rate compared to MC-oracle. This is an accepted result from active learning (although our experiments with 1NN and NB show no difference between the two on the VLBA data set).

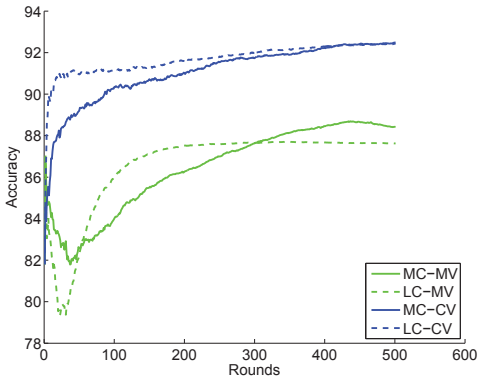
An open research question is whether the LC selection strategy can also benefit the low-cost labeling strategies MV and CV. Figure 4 shows preliminary results using LOG that demonstrate the expected results from active learning. Namely, LC selection leads to faster gains in classifier accuracy than the MC strategy. This is in spite of both LC-MV and LC-CV having poorer training set reliability compared to MC-MV and MC-CV. However, we have also observed cases where MC either outperforms or is on par with LC in terms of classifier accuracy. In future work, we will investigate the conditions under which LC outperforms MC with low-cost labeling.

5. Collaborative Learning

The family of learning techniques (MC-MV, LC-MV, MC-CV, and LC-CV) presented in this paper are part of a new learning framework, called *collaborative learning*, that we are developing for learning in multi-view, distributed environments. Our research hypothesizes that if individual learners can query each other for labels and share newly labeled examples, they can generate high-quality training data that quickly improves classifier performance at the node level. The key components of collaborative learning are example selection, label unification, and example broadcast. Figure 5 depicts a single round of collaborative learning implemented on a fully-connected network with four nodes. Each learner L_i has its view of the labeled ($\mathcal{L}_i$)



(a) Reliability, Logistic



(b) Accuracy, Logistic

Figure 4. LC (dashed lines) example selection learns more quickly than MC (solid lines) example selection in terms of training set reliability (top) and test set accuracy (bottom) using LOG.

and unlabeled ($\mathcal{U}_i$) data. Here, learner 3 queries its neighbors and receives labels l_1, l_2, l_4 from them.

Our approach seeks to classify incoming data for distributed sensor networks that naturally produce multi-view data. Each node of a sensor network observes the same phenomena from a different vantage point. The goal is to make reliable decisions *in situ*, enabling fast responses to local events and conditions. Real-world situations in which this technology is needed include classifying volcano activity using a distributed seismic network, monitoring vehicle traffic in a highway system, or tracking people in an airport to detect suspicious activity. In each case, decision-making at the node level can trigger responses such as simultaneous observation by an attached camera, activation of high-cost sensors, or increasing the data collection rate.

6. Conclusions and Future Work

This paper highlights research issues that arise from combining different aspects of co-training, ensemble

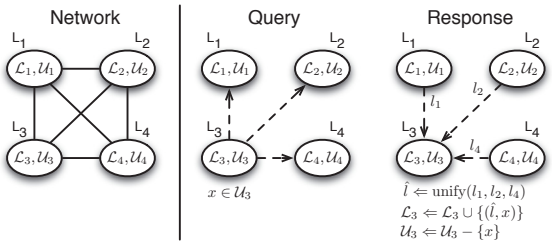


Figure 5. A single round of collaborative learning.

learning, and active learning. Because our work on collaborative learning is focused on the generation of reliable training data from low-cost multi-view labelers, we are interested in labeling strategies that minimize the injection of label noise, improve classification accuracy relative to self-labeling, and enable the use of low-confidence example selection in order to more efficiently improve classifier performance.

There are several open questions and avenues for further improvement in distributed, collaborative learning learning. Minimizing label noise is a central issue. We are currently investigating the impact of different learning configurations and parameters such as choice of base learner, number of initial labeled examples, network size, strategies for labeling and example selection, and other abstention policies, to determine their effects on training data reliability and classifier performance.

Acknowledgments

The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. This work made use of the Swinburne University of Technology software correlator, developed as part of the Australian Major National Research Facilities Programme and operated under license. We gratefully acknowledge the help of David Thompson and Walid Majid in providing the VLBA data used for this experiment. We thank Terran Lane for his feedback. This work is supported by NSF grant IIS-070568 and was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

References

Atla, A., Tada, R., Sheng, V., and Singireddy, N. Sensitivity of different machine learning algorithms to noise. *Journal of Computing*, 26(5):96–103, 2011.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proc. of the Conf. on Comp. Learning Theory*, pp. 92–100, 1998.

Brodley, C. E. and Friedl, M. A. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

Chapelle, O., Schölkopf, B., and Zien, A. *Semi-Supervised Learning*. The MIT Press, 1 edition, March 2010. ISBN 9780262514125.

Cohn, D., Atlas, L., and Ladner, R. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

Dietterich, T. G. Ensemble methods in machine learning. *Lecture Notes in Comp. Sci.*, 1857:1–15, 2000.

Kalapanidas, E., Avouris, N., Craciun, M., and Neagu, D. Machine learning algorithms: A study on noise sensitivity. In *Proc. of the 1st Balcan Conf. in Informatics*, pp. 356–365, 2003.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *Proc. of the ACM Conf. on Research and Development in Information Retrieval*, pp. 3–12, 1994.

Muslea, I., Minton, S., and Knoblock, C. A. Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27:203–233, 2006.

Napier, P.J., Bagri, D.S., Clark, B.G., Rogers, A.E.E., Romney, J.D., Thompson, A.R., and Walker, R.C. The very long baseline array. *Proceedings of the IEEE*, 82(5):658–672, 1994.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proc. of the 22nd International Conf. on Machine learning*, pp. 625–632, New York, NY, USA, 2005. ACM.

Pechenizkiy, M., Tsymbal, A., Puuronen, S., and pechenizkiy, O. Class noise and supervised learning in medical domains: The effect of feature extraction. In *Proc. of the 19th IEEE Symposium on Computer-Based Medical Systems*, pp. 708–713, 2006.

Pierce, D. and Cardie, C. Limitations of co-training for natural language learning from large datasets. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2001.

Wolpert, D. H. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8 (7):1341–1390, 1996.